A Dynamic "Fixed Effects" Model for Heterogeneous Panel Data

preliminary draft, comments welcome

Diana Weinhold♦
London School of Economics

April 1999

abstract: This paper introduces a dynamic panel data model in which the intercepts and the coefficients on the lagged endogenous variables are specific to the cross section units, while the coefficients on the exogenous variables are assumed to be normally distributed across the cross section. Thus the model includes mixture of fixed coefficients and random coefficients, which I call the "MFR" model. The paper shows that this model has several desirable characteristics. In particular, the model allows for a considerable degree of heterogeneity across the cross section both in the dynamics and in the relationship between the independent and dependent variables. Estimation of the MFR model produces an estimate of the variance of the coefficients across the cross section units which can be used as a diagnostic tool to judge how widespread a relationship is and whether pooling of the data is appropriate. In addition, unlike LSDV estimation of dynamic panel models, the MFR model does not produce severely biased estimates when T is small.

Key words: dynamics, panel data

♦ Development Studies Institute, London School of Economics, Houghton Street, London WC2A 2AE. Email: D.Weinhold@lse.ac.uk

This research was conducted while the author was in residence at the Center for Development Research (ZEF) at the University of Bonn, Germany. The author is gratefully indebted to Clive W.J. Granger for his ongoing guidance and support. In addition the author would like to acknowledge the useful comments and encouragement Andrew Levin and of all the participants in Valerie Ramey's and Haly Edison's group at the 1998 COFFEE workshop in Chicago, IL. All errors and omissions are my own.

1. Introduction

Panel data models in Macroeconomics have become increasingly popular in the past decade with the increased availability of cross country data sets that span 20 years or more. There are several key advantages of using panel data over a single time series or cross section data set. In cases where there is limited time series data available for each country there may be insufficient power of tests of hypotheses. If it is possible to impose some homogeneity conditions upon the parameter across countries then a panel data model will afford additional power and may allow the detection of relationships not apparent from the individual time series. Unlike cross section models, with a panel model it is possible to control for the country-specific, time invariant characteristics through the use of country-specific intercepts or "fixed effects." In addition, many processes display dynamic adjustment over time and ignoring the dynamic aspect of the data is not only a loss of potentially important information, but can lead to serious misspecification biases in the estimation. Including lagged dependent variables in a model can also control to a large extent for many omitted variables.

In practice there are two important econometric problems in estimating dynamic panel data models. The first is that parameters estimates are known to be biased in models with fixed effects and lagged dependent variables, and the second is that the homogeneity assumptions that are often imposed on the coefficients of the lagged dependent variable can lead to serious biases when in fact the dynamics are heterogeneous across the cross section units. This paper proposes a dynamic, fixed effects panel data model which reduces both of these problems. The estimation of the model does not require instrumental variables and the model has the additional benefit of providing the researcher with diagnostic information about the extent of heterogeneity in the panel.

The paper proceeds as follows. In section 2 the econometric issues surrounding the estimation of dynamic panel models are discussed, and some current estimation techniques are briefly reviewed. In section 3 I introduce an alternative model specification (the "MFR" model) which addresses the problems discussed in section 2 and discuss its properties. Section 4 presents the results of Monte Carlo simulation and a simple empirical exercise which illustrate the advantages of the MFR model. Section 5 summarizes the paper and concludes. The estimation algorithm for the MFR model is presented in the appendix.

2. Brief Discussion of Econometric Estimation of Dynamic Panel Data Models

Consider the simple model,

$$y_{it} = \gamma y_{it-1} + \beta x_{it} + e_{it}$$

where $e_{it} = \mu_i + \nu_{it}$ and $i = 1, \ldots, N$ cross section units and $t = 1, \ldots, T$ time periods. There is a clear simultaneity problem as the lagged dependent variable y_{it-1} is correlated with the error term e_{it} by virtue of its correlation with the time-invariant component of the error term, μ_i. Nickell (1981) has shown that even if the "fixed effects" (FE) or Least Squares Dummy Variable (LSDV) is used, y_{it-1} will still be correlated with the error term and the resulting bias will be of $O(1/T)$. Andersen and Hsiao (1981) and Hsiao (1986) both provide extensive discussions of this bias.

The usual approach for dealing with this problem is to first first-difference the data to remove the μ_i which yields:

$$y_{it} - y_{it-1} = \gamma (y_{it} - y_{it-1}) + \beta (x_{it} - x_{it-1}) + (e_{it} - e_{it-1})$$

Then, because Δy_{it-1} is correlated with the first difference error term it is necessary to instrument for it. Andersen and Hsiao (1981) have suggested using Δy_{it-2} or y_{it-2} as an instrument as these terms are not correlated with $\Delta e_{it} = \nu_{it} - \nu_{it-1}$. Arellano (1989) showed that an estimator that uses the levels for instruments has no singularities and displays much smaller variances than does the analogous estimator that uses differences as estimators. Holtz-Eakin et. al.(1988) adopt the approach to panel VAR's in a framework for testing Granger causality in panels and suggest using a time-varying set of instruments that includes both differences and levels. In addition other instruments have been suggested by a succession of researchers[1]. In practice however it is often difficult to find good instruments for the first-differenced lagged dependent variable, which can itself create problems for the estimation. Kiviet (1995) shows that panel data models that use instrumental variable estimation often lead to poor finite sample efficiency and bias. Using a broad array of Monte Carlo simulations he finds,

> In particular situations it seems that valid orthogonality restrictions can better not be employed when composing a set of instrumental variables. It is difficult to find clues on when which instrumental variables are better put aside in order to avoid serious small sample bias or relatively large standard deviations, which both entail poor estimator efficiency. As yet, no technique is available that has shown uniform superiority in finite samples over a wide range of relevant situations as far as the true parameter values

[1] For example, see Arellano and Bond(1991), Keane and Runkle (1992), Arellano and Bover(1995), and Ahn and Schmidt(1995). Baltagi (1995) provides a useful overview of all these papers.

and the further properties of the data generating mechanism are concerned. (p.72)

Fortunately, as *T* gets larger this bias becomes less of a problem. Nevertheless it would be useful to have an estimator which did not have such a large bias for small *T* and which did not require instrumental variables estimation. In particular, many cross country data sets have time dimensions of between 15 to 25 years, at which point it is hard to judge whether the Nickell bias or a weak instrument set will do more harm to the estimation.

An additional problem of introducing dynamics into a panel data model is the potential bias induced by heterogeneity of the cross section units. Pesaran and Smith (1995) have explored this problem in depth. They show that parameter estimates derived from pooled data are not consistent in dynamic models even for large *N* and *T*. In particular consider a model in which the coefficient on the lagged dependent variable is constrained to be equal across all cross section units so that we have:

$$y_{it} = a_i + gy_{it-1} + bx_{it} + e_{it}$$

there could be significant bias introduced if in fact the coefficients on the lagged dependent variable are not constant across the cross section. In this case the difference between the actual value and the estimated coefficient times the dependent variable, $(g_i - \bar{g})y_{it-1}$, will be a component of the error term[2] and this serial correlation induces bias and inconsistency in the estimation. As Pesaran and Smith (1995) point out, this bias is distinct from the fixed effects Nickell bias discussed above and cannot be addressed with instrumental variables estimation. In fact, in the case of macroeconomic cross country dynamic models the Pesaran and Smith bias could be much more severe than the former considering the larger *T* dimension of these data sets and the likelihood of cross country heterogeneity.

3. An Alternative Dynamic, "Fixed Effects" Panel Data Model

In this paper I propose an alternative specification for dynamic panel data models which allows for greater heterogeneity in the parameters than do the traditional models and has the additional benefit of producing considerably less biased parameter estimates than the fixed effects estimator with small *T*. This reduction in bias may make it more reasonable to forgo instrumental variables estimation in cases where the instruments are poor, a common problem with cross country panel data sets. In addition, the model provides its own internal diagnostics which can warn a researcher of excessive panel heterogeneity. In particular consider the specification:

$$y_{it} = a_i + g_i y_{it-1} + b_i x_{it} + e_{it}$$

where the coefficient on the lagged dependent variable is country-specific and the coefficient on the exogenous explanatory variable *x* is drawn from a random distribution

[2] Again, see Pesaran [19].

with mean $\bar{b}$, so that $b_i = \bar{b} + u_i$. I call this a "mixed fixed and random coefficients" (MFR) model after Hsiao (1989) who originally developed the estimation algorithm for a combination of fixed and random coefficients in a non-dynamic, non-fixed-effects panel data model of regional electricity demand.

It is important to explain why this particular combination of fixed, individual-specific coefficients on the lagged dependent variable and random coefficients only on the lagged independent variables has been chosen for the MFR model. First, note that it is inappropriate to simply introduce lagged dependent variables into a "random coefficients model". To illustrate, consider a simple model:

$$y_{it} = a_i + g_i y_{it-1} + b_i x_{it} + e_{it}$$

where $g_i = \bar{g} + h_{1i}$, $b_i = \bar{b} + h_{2i}$ and h is a random disturbance. The model can thus be rewritten as:

$$y_{it} = a_i + \bar{g} y_{it-1} + \bar{b} x_{it} + x_{it}$$

where $x_{it} = h_{1i} y_{it-1} + h_{2i} x_{it} + e_{it}$. There is clearly a serious problem with this specification as the error term is correlated with the lagged dependent variable. Moreover, the traditional econometric remedy for such models, using instrumental variable estimation, may not be possible in this situation. It is very difficult, if not impossible, to find instruments which are correlated with the LHS variables but not with the error term.[3] If we instead model the coefficient on the lagged dependent variable as fixed rather than random, but constrain it to be equal across all cross section units so that we have:

$$y_{it} = a_i + g y_{it-1} + b_i x_{it} + e_{it}$$

there could still be significant biases introduced if in fact the coefficients on the lagged dependent variable are not constant across the cross section, as discussed in section 2 of this paper. It is thus clear that if we suspect heterogeneity in the panel it is prudent to allow as much flexibility in the estimation of the coefficient on the lagged dependent variable as possible. However we cannot allow both the coefficients on the explanatory variables and on the lagged dependent variable to be random.

The MFR model avoids both of the problems outlined above. It allows for heterogeneity in both the coefficient on the lagged endogenous variable and on the exogenous variable without introducing the simultaneity problem. However, it has not yet been shown why this model should be less biased with small *T* than traditional fixed effects dynamic panel data models. Section 4 presents a series of Monte Carlo simulations that demonstrate that the bias in the estimates of the coefficient on the exogenous variable are quite small even when the time series is as short as $T = 5$. If it is desired, the MFR model can be estimated using any of the instrument sets suggested in the literature. However, given that the

[3] For an explicit treatment of this problem see Pesaran [19] p. 24.

instruments are often of quite poor quality, researchers using cross country panel data which tend to have time dimensions of about 15 to 30 years may find that using a MFR model is preferable to traditional models. In addition, unlike traditional panel models, the MFR model provides panel diagnostics which are discussed in the next section.

3.1 Heterogeneity and the MFR Model as a Diagnostic Tool

In addition to the properties discussed above, the MFR model estimation provides an estimate of not only the mean, but also the variance of the random coefficients. The estimated variance of b, which we denote s_b^2, can be used as a diagnostic tool to determine the extent of heterogeneity in the relationship in question. If the estimated variance is quite large relative to the coefficient estimates, this is a signal of significant heterogeneity in the panel. Rather than continuing with the panel data analysis at that point it may be prudent to rather investigate the heterogeneity in greater detail to assess the appropriateness of the specification and/or the pooling of the data. However, if the variance is quite small then a researcher can have additional confidence that the conclusions of the estimation are quite general across the panel. In a complementary paper (Weinhold (1998)) I propose a "rule of thumb" for interpreting the estimated variance from the MFR model in the context of causality testing. In particular I suggest multiplying the standard error of the mean estimate by $\sqrt{N}$ and constructing a confidence interval around zero that is twice this distance on either side. The area that falls within this interval is interpreted to correspond to observations that are not significantly different from zero. Thus,

$$\text{Prob}(x = 0) = \text{Prob}\left(0 - 2\sqrt{N}\,\text{se}_{\hat{b}} \le x \le 0 + 2\sqrt{N}\,\text{se}_{\hat{b}}\right)$$

Another use of the estimated distribution from the MFR model is to recover estimates of the individual cross section coefficients in the case in which the time dimension is too short to have confidence in the individual time-series estimates[4]. Specifically, under the assumption of normality we have $b_{1i} \sim N(\bar{b}_1, s_b^2)$. Although the cross section time series are too short for reliable estimates of the individual b_{1i}'s, these estimates can be used to estimate the *order* of the parameter values across cross section units. We can denote these ordered estimates as b(j) with b($j+1$) > b(j) for all j[5]. Then, for example, b_i = b(j) implies that b_i is the jth largest estimate. If $E_{N(J)}$ is the jth normal score from a sample of size N, we can construct a new estimate $\tilde{b}_i = sE_N(j) + \bar{b}$. That is, by scaling the normal score using the mean and standard deviation estimated in the first two steps we derive the new estimates $\tilde{b}_i$'s. These estimates will have the correct mean and variance and will be perfectly normally distributed. Furthermore, they should be more reliable than those estimated directly from a single series if the assumption that they are randomly distributed is valid.

[4] This approach was originally suggested by Prof. C.W.J. Granger.
[5] Ties can be broken by averaging equal values or by random allocation.

4. Monte Carlo Simulations and a Simple Empirical Example

4.1 Monte Carlo Trials

Monte Carlo simulation trials were run to gauge the extent of any bias in the MFR estimates for different magnitudes of *T* and *N*. In particular data were generated that I feel approximates the nature of many cross country data sets, with cross-section specific fixed effects and individual dynamics. Thus panels of different dimensions were generated from the model[6]:

$$x_{it} = \alpha_{1i} + 0.41x_{it-1} + \varepsilon_{1it}$$
$$y_{it} = \alpha_{2i} + \gamma_i y_{it-1} + \beta_i x_{it} + \varepsilon_{2it}$$

where $\alpha_{1i} \sim RNDU[0.2,\ 1.2]$, $\alpha_{2i} \sim RNDU[0.5,\ 3.5]$, $\gamma_i \sim RNDU[-0.003,\ 0.997]$, and $\beta_i \sim N(0.2,\ 1.0)$. Both errors terms were drawn from the standard normal distribution, as were the initial values of both *x* and *y*. The model was allowed to iterate 30 times before the panel data of *x*'s and *y*'s were used for estimation.

Kiviet (1995) performed a wide variety of Monte Carlo simulations for different dynamic panel data estimators. He shows that although it is biased, the LSDV has a relatively small standard deviation compared to the asymptotically consistent IV and GMM estimators. He therefore proposes an estimator in which an approximation to the Nickell bias is calculated and the LSDV estimator is directly corrected, and finds this estimator to be more efficient than the alternatives. However, the Kiviet estimator is derived under the assumption of homogenous parameter values across the cross section, as are the alternative IV and GMM estimators. We can nevertheless use his estimates of the biases as a general point of comparison, taking into account the fundamental differences in model design. Since the actual LSDV Nickell bias depends on the value of γ, the coefficient on the lagged dependent variable, I do not expect to obtain the normal Monte Carlo LSDV biases that are generated under homogenous parameters. Rather, the bias on the LSDV estimator reported here will be a combination of the Nickell bias *and* the Pesaran and Smith bias. Since many real world applications are susceptible to both sources of bias (and I have argued that in many cases the Pesaran and Smith bias might be the more damaging of the two) I feel that it is interesting to observe their joint impact on the standard LSDV estimator.

Panel dimensions that closely resemble the dimensions of many cross country data sets were used. In particular I examine model performance for T= 5, 10, 25 and 50 years for panels of N= 20, 40 and 80 cross section units. For each simulation both the MFR model as well as an LSDV model were estimated using the generated panel data. This was

[6] Some simulations were also run constraining the coefficient on the lagged dependent variable to be constant across countries. If anything the reduction in bias from the MFR model was even greater in these cases. However, computation time is quite long for the larger panels and I feel that heterogeneous dynamics is more likely representative of cross-country data sets. Thus heterogeneous dynamics were used for the whole simulation presented here.

repeated 10,000 times and the average estimate of $\bar{b}$ from each model is presented in table 1. Since we know the true value of $\bar{b}$ is 0.20 we can observe that the biases of the models directly and compare their magnitudes. Table 2 presents the estimates of the variance of b_i, s_b^2. Again, since we know that s_b^2=1.0 we can readily observe any bias from the MFR estimation. Finally, since I am allowing country-specific dynamics in the model (i.e. g_i is uniformly distributed across the cross section) there is no single "true" parameter value g in the model.

The results from table 1 illustrates that, unlike the traditional LSDV estimator, the mean estimate of b is not severely biased by a short time dimension in the data for the typical cross section dimensions found in cross country data sets. On average the biases range between 0.001 and 0.003. There is a slightly larger upwards bias of 0.0105 on the estimate when *N*=80 and *T* =5 however. Unfortunately, this dimensionality is the most comparable to the Kiviet (1995) results, although Kiviet's models have one more time observation which may make a big difference. To check this I rerun the simulation for *T*=6 and *N*=100, emulating Kiviet as closely as possible by running 1000 replications. Our average estimate of the mean is .2021 (the LSDV estimate in this case is 0.1832) The calculated bias, shown at the end of the second column in table 3, is thus 0.0021 which compares quite favorably with the alternative estimators. In addition, table 2 presents the standard deviation of the estimates which show the clear pattern of root-N consistency.

For panels with time dimensions of 10 or greater, the average estimate of the MFR model seems to hover between 0.2012 and 0.2025 without showing much systematic evolution related to the time or cross section dimensions. On the other hand I find that the bias on the LSDV is significant in all of the panels and actually increases as T grows larger[7]. This confirms my earlier suspicion that in panels with heterogeneous dynamics, a longer time dimension does not necessarily justify the use of the LSDV estimator.

As table 4 illustrates, while the average estimated mean of b in the MFR model appears fairly invariant to the time dimension, the estimate of the variance, s_b^2, is highly biased upwards for very small *T*. In particular we can observe that the estimated variance for *T* = 5 is between 7 and 10 times the actual value. This upward bias initially declines sharply with the number of time periods so that by *T* = 10 the estimated variance is 1.24 compared to the actual value of 1.0. The bias declines at a decreasing rate after that; at *T* = 25 the estimated variance has only a very small bias of 1.07, and at *T*= 50 it has decreased to 1.03. Thus researchers using the MFR model with short time series should take into account that the variance estimate is highly biased upwards. This would effect the diagnostic ability of the model if the bias is not taken into account, but the simulations

[7] I assume this is due to the use of heterogeneous dynamics in the simulated data and is the result of the Pesaran and Smith (1995) bias. One possibility is that the two biases somehow mitigate each other when T is small, but that as T grows larger and the Nickell bias shrinks, the Pesaran and Smith bias come to dominate. In this case the Pesaran and Smith Bias is clearly negative.

have shown that the mean estimate is much less biased the LSDV estimator which could make the model a good choice regardless.

Finally, in order to more directly compare the bias of the MFR model with those of alternative estimators under heterogeneous dynamics a final set of Monte Carlo simulations is run. In particular it is informative to investigate the properties of some of the more common dynamic panel data estimators, which have been derived under homogenous dynamics assumptions, when in fact the dynamics are generated to be unit-specific.

Thus, following Judson and Owen (1997) the simulation is expanded to include a simple Anderson and Hsiao (1981) estimate as well as a GMM estimator proposed in Arellano and Bond (1991). Anderson and Hsiao first difference the model, yeilding $\Delta y_{i,t} = g\Delta y_{i,t-1} + b\Delta x_{i,t} + \Delta e_{it}$ where $\Delta y_{it} = y_{i,t} - y_{i,t-1}$ and so on. They then suggest estimating the model using the second lagged level of the dependent variable as an instrument for the the lagged first differenced dependent variable. Thus we have $\hat{d}_{AH} = (Z'X)^{-1}Z'Y$ where Z is a matrix of instruments, X is a matrix of regressors and Y is a vector of dependent variables.

A GMM estimator suggested in Arellano and Bond (1991), and explored by Kiviet (1995) as well as by Judson and Owen (1997), uses both lags of the dependent as well as exogenous regressors as instruments. Briefly, the estimator can be expressed as: $\hat{d}_{GMM} = (X'Z^* A_N Z^{*'} X)^{-1} X'Z^* A_N Z^{*'} Y$ where X and Y are defined as above, and Z_i^* is a block diagonal matrix[8]. Different choices for A_N result in different GMM estimators. Judson and Owen compare two alternatives and find that one of these produces both smaller biases and smaller standard deviations of the estimates, and thus we focus on this version alone. In particular their chosen estimator uses $A_N = \left(\frac{1}{N} \sum_i Z_i^{*'} H Z_i^* \right)^{-1}$ where H is a T-2 square matrix with twos in the main diagonals, minus ones in the first subdiagonals, and zeros otherwise. In the simulation presented here a restricted version of this GMM estimator is employed in which the number of instruments used is set to three[9].

While both of these estimators have been found to be consistent when the data generating process in each of the cross section units is homogenous except for an individual specific intercept term or "fixed effect," most empirical macro economic models use countries as the cross section unit of analysis. In these cases it is almost always the case that the dynamic properties, especially the short run dynamics, are quite distinct across countries. In addition the processes determining the evolution of the exogenous variables as well as

[8] For full details see Judson and Owen (1997). Essentially the sth block of Z_i^* is given by $((y_{i1} \ldots y_{is} x_{i1} \ldots x_{i,s+1})$ for s=1,…,T-2. Then $Z^* = (Z^*_1, \ldots, Z^*_N)'$.

[9] For the T=10 case I also experimented with a GMM estimator with 5 instruments, with no significant change in the results.

the salient omitted variables may well vary quite a bit from country to country, so that the restrictive assumptions of the theoretically consistent estimators may start to diverge alarmingly far from reality. For this Monte Carlo simulation the data was generated in a similar manner to the simulations from Table 1, with the small difference that the random numbers were generated in this case using seeds and stored in advance in a matrix. Since the GMM estimators are originally designed for use when N is relatively large and T relatively small (and computational demands increase considerably with higher T) comparisons were done only in the T=5 and T=10 cases. In addition, due to the extra computational requirements of the GMM estimator the number of replications in these simulations was restricted to either 4000 or 2000 replications, depending on the T and N dimensions, and is so noted in table 5.

The results, in table 5, yeild similar bias estimates for the MFR and LSDV estimation as in table 1, as one would expect. However both the instrumental variables esimators perform very poorly under the more realistic heterogeneous dynamics. The Anderson and Hsiao estimator shows no clear pattern in terms of the sign of the bias, but clearly has a high variance around the true mean value of 0.2. The GMM estimator, on the other hand, displays a persistent and highly significant negative bias. The only conclusion of such an exercise is that while instrumental variables estimators may still be the best choice for microeconomic data that more closely adheres to the homogeneity assumptions under which these estimators gain consistency, it is not so clear that these are the superior choice for models of macro-economic processes across countries or other cross section units in which the dynamics are likely to be heterogeneous.

4.2 A Simple Empirical Example

In order to illustrate empirically the advantages of the MFR model over traditional models I consider a simple dynamic model of the impact of the share of GDP of aid (AIDSH), foreign direct investment (FDISH), and gross domestic investment (GDISH) on the growth (GGDP) in a panel of 32 less developed countries from 1976 to 1995[10].

$$GGDP_{it} = a_i + g_1 GGDP_{it-1} + g_2 GGDP_{it-2} + b_{11} GDISH_{it-1} + b_{12} GDISH_{it-2} + b_{21} FDISH_{it-1} + b_{22} FDISH_{it-2} + b_{31} AIDSH_{it-1} + b_{32} AIDSH_{it-2} + e_{it}$$

This model was estimated following the Holtz-Eakin et.al. (1988) methodology for causality testing in panels, in which the model is first-differenced to eliminate the fixed effects and the first differenced lagged dependent variable is instrumented for with a mix of lagged differences and levels of the (several times lagged) dependent variable as described in section 2. The model was then modified to be estimated as a MFR model by specifying fixed, country specific coefficients on the lagged dependent variables and random coefficients on the lagged explanatory variables so that it became:

[10] This example is taken from Nair and Weinhold (1998).

$$GGDP_{it} = a_i + g_{i1}GGDP_{it-1} + g_{2i}GGDP_{it-2} + b_{11i}GDISH_{it-1} + b_{12i}GDISH_{it-2} + b_{21i}FDISH_{it-1} + b_{22i}FDISH_{it-2} + b_{31i}AIDSH_{it-1} + b_{32i}AIDSH_{it-2} + e_{it}$$

where $b_{jki} \sim N(\bar{b}_{jk}, s^2_{jk})$. The lagged exogenous variables were orthogonalized with respect to the lagged dependent variables and each other so that the coefficients would be independent and the variance estimates would not be influenced by each other. Since the time span is 20 years I elect not to make any adjustments to the estimated variances, although it is expected that the estimates will be slightly biased upwards. Table 4 summarizes the essential results from the estimation of both the MFR model and the Holtz-Eakin model[11]. In particular, for the MFR model the table shows the estimated mean, variance and standard deviation, and for the Holtz Eakin model it presents the coefficient estimates and the results of an F-test for the joint significance of the two lags of the given variable.

The results from the Holtz-Eakin estimation in table 6 seem to provide strong evidence that both GDISH and AIDSH Granger-cause GGDP. However, the MFR results clearly show that there is enormous variability across countries. The MFR estimation results find no statistical evidence of Granger causality in the panel. In fact, when individual country-by-country regressions are run I find that in fact there is huge heterogeneity in each of these relationships, both in terms of magnitude, sign and statistical level of significance. Although the individual country regressions are not particularly reliable themselves due to the limited time dimension, it is clear that the strong conclusions drawn by the Holtz-Eakin estimation are misleading. Furthermore, as is illustrated in table 7 where the number of countries has been reduced to *N*=28, if the set of countries is varied and the models re-estimated, the estimated coefficients and their statistical significance are susceptible to dramatic changes in the Holtz-Eakin model. The mean estimates are more stable in the MFR model, with the estimated variances adjusting more dramatically and providing some guidance in interpreting the results in the context of how widespread the relationship in question is in the panel. (See Nair and Weinhold 1998, from which this example was drawn, for a comprehensive discussion of heterogeneity and causality modeling using this data set).

5. Conclusions

As many economists are beginning to work with cross-country panel data sets there is an increasing interest in panel data estimation. In this paper I have promoted the use of a "mixed fixed and random" coefficients model, or MFR model, for cases in which a researcher would like to include both "fixed effects" and lagged dependent variables in a panel data model. This paper has shown that the MFR model allows for much more flexibility of the coefficients across the cross section than do traditional models. This property allows the MFR model to avoid the simultaneity biases that arise from constraining the coefficient on the lagged dependent variable to be constant across countries and provides the researcher with a diagnostic tool to judge how widespread a

[11] For a comprehensive analysis of these models and related issues see Nair and Weinhold (1998).

particular relationship is across the panel. This is especially important for cross-country data sets in which there can be considerable heterogeneity. In addition, simulations have shown that the estimates of the mean coefficient value on the exogenous variables in the MFR model have a much smaller bias than do estimates from LSDV estimation. This can be a major advantage when the instrument set necessary for a Andersen-Hsiao type solution to the LSDV bias is of poor quality. In these cases it may be preferable to use an MFR model without IV estimation.

6. Tables

Table 1: Monte Carlo Comparison of MFR and LSDV Estimation

	N = 20		N = 40		N = 80	
	LSDV	**MFR**	**LSDV**	**MFR**	**LSDV**	**MFR**
T = 5	0.1807	0.2049	0.1802	0.2014	0.1847	0.2105
T = 10	0.1851	0.1998	0.1860	0.2012	0.1885	0.2026
T = 25	0.1812	0.2027	0.1782	0.2024	0.1777	0.2024
T = 50	0.1753	0.2023	0.1715	0.2012	0.1670	0.1991

Note: # reps = 10,000 $\bar{b}$ **=0.2**

Table 2: Standard deviation of estimate of beta (shows root-N consistency)

	N = 20		N = 40	N = 80	
	LSDV	**MFR**	**MFR**	**MFR**	**LSDV**
T = 5	0.30715	0.43178	0.30091	0.20752	0.15699
T = 10	0.25472	0.24886	0.17728	0.12518	0.12893
T = 25		0.23134	0.16448	0.11575	
T = 50	0.20571	0.22876	0.16124	0.11315	0.09879

Note: # reps = 10,000 $\bar{b}$ **=0.2**

Table 3: Kiviet (1995) Monte Carlo Bias Estimates Compared with MFR Model

Estimator	Bias	Estimator	Bias
GMM1	0.000	IV Δ X	0.053
GMM2	-0.006	OLS	-0.068
AH Δ	0.023	OLS_t	-0.067
AHX Δ	0.007	LSDV	0.005
AHL	-0.001	$LSDV_b$	0.007
AHXL	0.176	$LSDV_c$	-0.011
IVAX	0.006	**MFR**	**0.002**

***b* =0.2, T=6, N=100 (#reps = 1000)**

Table 4: Monte Carlo MFR Estimates of s_b^2

	N = 20	**N = 40**	**N = 80**
T = 5	7.663	10.047	9.392
T = 10	1.242	1.243	1.247
T = 25	1.069	1.065	1.068
T = 50	1.030	1.031	1.035

Note: # reps = 10,000 s_b^2 =1.0

Table 5: Monte Carlo Comparison of MFR, LSDV, AH and GMM Estimation

	N = 20				N = 40				N = 80			
	MFR	LSDV	AH	GMM	MFR	LSDV	AH	GMM	MFR	LSDV	AH	GMM
T=5	.2082 [a]	.1867 [a]	.2145 [a]	.1688 [a]	.2098 [a]	.1809 [a]	.1547 [a]	.1497 [a]	.2047 [b]	.1813 [b]	.2300 [b]	.1427 [b]
T=10	.2008 [a]	.1869 [a]	.2196 [a]	.1723 [a]	.2033 [b]	.1931 [b]	.2462 [b]	.1777 [b]	.2057 [b]	.1923 [b]	.1775 [b]	.1751 [b]

a: # reps = 4,000 $\bar{b}$ =0.2

b: # reps = 2,000

Table 6: Dependent variable = Growth (GDP), N=32, T=20

	MFR Estimation			Holtz Eakin et. al. Estimation		
Variable	Mean	Variance	St. Error	coeff. est.	Joint F-stat	p-value (F)
GDISH_1	-4.025	4479.723	17.911	-21.465	9.23	0.0012
GDISH_2	-18.613	5186.013	23.842	-5.465		
FDISH_1	0.492	42.900	0.924	0.205	1.08	0.3400
FDISH_2	0.211	42.367	1.000	-0.049		
AIDSH_1	0.114	49.597	0.274	0.683	48.70	>.0001
AIDSH_2	0.858	149.100	0.559	0.109		

Table 7: Dependent variable = Growth (GDP), N=28, T=20

	MFR Estimation			Holtz Eakin et. al. Estimation		
Variable	Mean	Variance	St. Error	coeff. est.	Joint F-stat	p-value (F)
GDISH_1	-3.872	4955.860	13.249	-18.385	6.949	0.0011
GDISH_2	-15.704	3067.529	17.731	-4.933		
FDISH_1	0.468	48.939	0.761	0.330	1.607	0.2017
FDISH_2	0.266	48.319	0.874	-0.0802		
AIDSH_1	-0.073	31.203	0.336	0.237	2.054	0.1297
AIDSH_2	1.053	166.268	0.534	0.116		

Appendix

A.1 Hsiao's Mixed Fixed and Random Estimation Algorithm

The following algorithm is available in GAUSS code via Email from the author upon request.

The algorithms presented here are from Hsiao [12] . The estimate of the mean of the random causal variable is:

$$\bar{b} = \left[\sum_{i=1}^{N} X_i' f_i^{-1} X_i - \sum_{i=1}^{N} X_i' f_i^{-1} Z_i \left(Z_i' f_i^{-1} Z_i\right)^{-1} Z_i' f_i^{-1} X_i\right]^{-1} \times \left[\sum_{i=1}^{N} X_i' f_i^{-1} Y_i - \sum_{i=1}^{N} X_i' f_i^{-1} Z_i \left(Z_i' f_i^{-1} Z_i\right)^{-1} Z_i' f_i^{-1} Y_i\right]$$

where $f_i = X_i \Delta_r X_i' + s_i^2 I_T$, X_i denotes the T time series observations of the designated random variables, Z_i denotes the observation of the variables with "fixed" coefficients, and s_i^2 is the estimated variance of the individual OLS regression of Y_i on X_i and Z_i. In addition,

$$\Delta = \frac{1}{N-1}\sum_{i=1}^{N}\left(b_i - N^{-1}\sum_{i=1}^{N} b_i\right)\left(b_i - N^{-1}\sum_{i=1}^{N} b_i\right)'$$

where b_i are the estimates of the coefficients on the X_i from an OLS estimation of Y_i on X_i and Z_i.

Then, Δ_r denotes the submatrix of Δ corresponding to the X_i. Thus, it is possible to read off directly from this matrix the estimates of the variance of the random coefficients. The fixed coefficient estimates may be recovered from:

$$b_{fi} = \left(Z_i' Z_i\right)^{-1}\left[Z_i'\left(Y_i - X_i \bar{b}\right)\right]$$

References

Ahn, S.C. and P. Schmidt, 1995. "Efficient estimation of models for dynamic panel data" *Journal of Econometrics* vol. 68

Anderson, T.W. and Cheng Hsiao, 1982. "Formulation and Estimation of Dynamic Models Using Panel Data", *Journal of Econometrics*, vol. 18.

Arrellano, M. and O. Bover, 1995. "Another look at the instrumental variables estimation of error-component models" *Journal of Econometrics* vol. 68

Arrellano, M and S. Bond, 1991. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations" *Review of Economic Studies*, 58

Baltagi, Badi H. 1995. Econometric Analysis of Panel Data, John Wiley & Sons Ltd. West Sussex, England

Holtz-Eakin, D. and W. Newey and H. Rosen, 1988. "Estimating Vector Autoregressions with Panel Data", *Econometrica* vol. 56, no. 6.

Hsiao, Cheng, Analysis of Panel Data, 1986. Cambridge University Press, Cambridge, England.

Hsiao, Cheng., 1989. "Modeling Ontario Regional Electricity System Demand Using a Mixed Fixed and Random Coefficients Approach," *Regional Science and Urban Economics*, vol. 19.

Judson, Ruth A. and Ann L. Owen, 1997. "Estimating Dynamic Panel Data Models: A Practical Guide for Macroeconomists" Finance and Economics Discussion Paper 1997-3, Federal Reserve Board, Washington D.C.

Keane, M.P. and D.E. Runkle, 1992. "On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous" *Journal of Business and Economics Statistics*, vol. 10.

Kiviet, Jan F. 1995. "On bias, inconsistency, and efficiency of various estimators in dynamic panel data models" *Journal of Econometrics* vol. 68.

Matyas, Laszlo and Patrick Sevestre, editors, 1992. Econometrics of Panel Data: Theory and Applications, Kluwer Academic Publishers.

Nair, U. and D. Weinhold, 1998. "Economic growth, openness to trade and FDI in less developed countries" unpublished manuscript

References (cont.)

Nickell, Stephen “Biases in Dynamic Models with Fixed Effects,” 1981. *Econometrica*, vol. 49, no. 6.

Pesaran, Hashem M. and R. Smith, 1995. “Estimating Long-Run Relationships From Dynamic Heterogeneous Panels”, *Journal of Econometrics* vol. 68.

Weinhold, D. 1996. "Investment, growth and causality testing in panels" (in French) *Economie et Prevision*, no. 126-5.

Weinhold, D. 1998. “Testing for causality in heterogeneous panel data: applications to cross-country growth analysis” unpublished manuscript (available by Email from the author upon request)